3.4 Sequence-based Data

Sequence-based data includes:

- Trace files
- FASTA files
- NCBI Trace ID relationships (*i.e.* relationship of trace files to sample-aliquot IDs, and genes)
- NCBI Short Read relationships
- Mutations (*e.g.* MAF)

About Aliquot Barcodes

The DCC requires the complete aliquot barcode (see "3.1 BCR Sample Identifiers - Aliquot Barcodes") so the entire TCGA enterprise including data provenance can be tracked and managed. Using aliquot barcodes currently involves extra steps for both data preparation and analysis. Those extra steps are:

1. Combining the plate and center IDs from the BCR plate sheet or Excel spreadsheet into a plate barcode (see BCR Plate Barcode) as a suffix to the analyte barcodes before you send your data to the DCC, and
2. Removing the plate and center ID, and possibly other parts of the barcode, from the aliquot barcode to aggregate data between different centers or data types.

Sheets | Charts | SmartArt Graphics | WordArt

96 Wellplate Sample Information Shipment QC File

Wellplate Barcode	0349	Institution:	Washington University
		Date of Shipment:	5/28/08

ROW	COLUMN	Biospecimen Barcode Side	Biospecimen Barcode Bottom	Tissue Type	Histology	Type DNA/RNA	Volume	Target Concentration
A	1		No Tube					
A	2		No Tube					
A	3		No Tube					
A	4		No Tube					
A	5	TCGA-12-0772-01A-01W	73443266	Brain	Glioblastoma multiforme	WGA DNA	400 uL	0.5 ug/uL
A	6	TCGA-06-0216-10A-01W	73443265	Blood	Normal	WGA DNA	400 uL	0.5 ug/uL
A	7		No Tube					

Figure 4 - Example of BCR plate sheet displaying plate barcode

The blue circle in Figure 4 shows an example of where to find the plate ID on the BCR plate sheet. The center ID can be obtained from Table 5. Aggregation of data using barcodes is discussed in Chapter 6 of the TCGA Data Primer. To mitigate the first step 1, the BCR has agreed to change the way they report the IDs to the GSCs/CGCCs so that the ID is always the complete aliquot barcode. The second step has to be completed regardless to compare results between data type (*e.g.* mutations verses gene expression).

All centers are required to submit IDs as aliquot barcodes. That includes Trace-sample relationship (TR) files, Mutation (MAF) files, and verbose coverage files (VCFs). Archives that fail validation will not be available for bulk (FTP/SFTP) download or *via* the Data Access Matrix (http://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) until the archive passes validation.

Trace Relationships

Trace files are submitted directly to the NCBI Trace Archive. GSCs are required to submit Trace Relationships to the DCC as Trace Relationship (TR) files. TR file consist of NCBI trace ID, the aliquot barcode associated with those trace file submissions, and the sequencing target ID (*i.e.* HUGO gene symbol or ConsReg). TR files are considered Level 1 data with a data type of "Trace-Sample Relationship" (see "2.2 Categorizing Data" in this document and/or "2.2 Categorizing Data" in the TCGA Data Primer). Please do not include actual trace or FASTA files in archives transferred to the DCC.

Trace relationship files should have the suffix "tr" and contain the prefix of the containing archive name (*e.g.* broad.mit.edu_GBM.ABI.1.tr). The data in that file should be *tab-delimited* with no leading spaces and modeled using the following ordered data elements as columns:

1. trace_id (*required*; NCBI Trace "ti")
2. aliquot_id (*required*; the BCR Aliquot Barcode; see "3.1 BCR Sample Identifiers - Aliquot Barcodes")
3. target_id (*will be required soon*; HUGO Gene Symbol or non-gene target ID; same as gene_name in NCBI Trace Archive)

That data allows the DCC to query NCBI Trace for additional metadata and relate this metadata to other experimental results by BCR aliquot barcode. As the gene_name (mapped to target_id here) field is not consistently populated in NCBI Trace, the DCC requires that the values be populated in the tr file.

Short-read Relationships

(*THIS SECTION, "SHORT-READ RELATIONSHIPS," IS NOT COMPLETE AND MAY NOT BE USED AT ALL)*

Short-read files are submitted directly to the NCBI Short Read Archive. GSCs are required to submit short-read relationships to the DCC as Short-read Relationship (SR) files. SR file consist of NCBI SRA, SRX, SRS, SRR IDs, the TCGA aliquot barcode associated with a short-read file submission, the submission date, load date, submission status, base counts, and the GSC's internal tracking ID. The SR file allows the DCC to query the NCBI Short Read Archive for additional metadata and relate this metadata to other experimental results by BCR aliquot barcode. SR files are considered Level 1 data with a data type of "Short Read-Sample Relationship" (see "2.2 Categorizing Data" in this document and/or "2.2 Categorizing Data" in the TCGA Data Primer). Please do not include actual short-read or BAM files in archives transferred to the DCC.

Short-read relationship files should have the suffix "sr" and contain the prefix of the containing archive name (*e.g.* broad.mit.edu_GBM.ABI.1.tr). The data in that file should be *tab-delimited* with no leading spaces and modeled using the following ordered data elements as columns:

1. aliquot_id (*required*; the BCR Aliquot Barcode; see "3.1 BCR Sample Identifiers - Aliquot Barcodes")
2. sra_acc (*required*; submission accession)
3. srx_acc (*required*; experiment accession)
4. srs_acc (*required*; sample accession)
5. srr_acc (*required*; run accession)
6. submission_date (*required*; date of run submission to Short Read Archive)
7. load_date (*required if available*; date the run was loaded into Short Read Archive)
8. submission_status (*required if available*; *i.e.* live, suppressed, hold till publish, killed)

9. base_count (*required;* number of bases contained in a run)
10. gsc_tracking_id (*required*; GSC internal tracking ID alias)

Values for columns 1-8 should be available directly from the NCBI Short Read Archive.

Mutation Data

Mutation annotation files should be transferred to the DCC. Those files should be formatted using the mutation annotation format (MAF) that is described below. The file names should have the suffix "maf" and contain the prefix of the containing archive name (*e.g.* broad.mit.edu_GBM.ABI.1.maf).

The following data are reported in MAF files:

Somatic mutations

- Missense and nonsense
- Splice site, defined as SNP within 2 bp of the splice junction
- Silent mutations
- Indels that overlap the coding region or splice site of a gene or the targeted region of a genetic element of interest.

SNPs

- Any germline SNP with validation status "unknown" is included.
- SNPs already validated in dbSNP are not included since they are unlikely to be involved in cancer.

Validation

All candidate somatic missense, nonsense, splice site and indels are retested by an independent (orthogonal) genotyping method. If the SNP is confirmed by an independent method, they are deemed valid. Silent mutations may be validated for the purpose of calculating the background mutation rate. No germline (SNP or indel) candidates are processed through validation. However, if the validation process reveals a given candidate somatic variation event to be germline or loss of heterozygosity, those validated data are reported in the validation file.

A *validated somatic mutation* is identified by (Verification_Status=Verified or Validation_Status=Valid) and Mutation_Status=Somatic.

MAF files have a data type of "Mutations". Putative (un-validated) somatic mutations or non-somatic mutations are considered Level 2 data and have controlled access only. Validated somatic mutations (defined above) are considered Level 3 data and open access.

Mutation Annotation Format File Fields

The format of a MAF file is tab-delimited columns. Those columns are described in Table 7 and are required in every MAF file. The order of the columns is important and will be validated by the DCC. Column headers and values are not case sensitive. Columns may allow null values (*i.e.* blank cells) and/or have enumerated values.

Table 7 - Mutation annotation format file column headers

Index	MAF Column Header	Description of Values	Null	Fixed
1	Hugo_Symbol	HUGO symbol for the gene, *e.g.* EGFR (HUGO symbols are *always* capitalized)	No	Set
2	Entrez_Gene_Id	Entrez gene ID, *e.g.* 1956	No	Set
3	GSC_Center	Genome sequencing center reporting the variant. One of hgsc.bcm.edu, broad.mit.edu, or genome.wustl.edu	No	Yes
4	NCBI_Build	NCBI human genome build number with decimal (e.g. 36.1, 36.2, etc.)	No	Set
5	Chromosome	chromosome number without "chr" prefix, e.g. X, 1, 2	No	Set
6	Start_position	mutation start coordinate (1-based coordinate system)	No	No
7	End_position	mutation end coordinate (inclusive, 1-based coordinate system)	No	No
8	Strand	one of "+" or "-".	No	Yes
9	Variant_Classification	one of Missense_Mutation, Nonsense_Mutation, Silent, Splice_Site_SNP, Frame_Shift_Ins, Frame_Shift_Del, In_Frame_Del, In_Frame_Ins or Splice_Site_Indel	No	Yes
10	Variant_Type	one of SNP, Ins or Del	No	Yes
11	Reference_Allele	the plus strand reference allele at this position	No	No
12	Tumor_Seq_Allele1	tumor sequencing (discovery) allele 1	No	No
13	Tumor_Seq_Allele2	tumor sequencing (discovery) allele 2	No	No
14	dbSNP_RS	dbSNP id (e.g. rs12345) or none or novel	No	Set
15	dbSNP_Val_Status	dbSNP validation status; one of byCluster, bySubmitter, byFrequency, by2hit2allele, byHapmap, none, or unknown	No	Yes
16	Tumor_Sample_Barcode	BCR Aliquot Barcode for tumor sample, i.e. TCGA-SiteID-PatientID-SampleID-PortionID-PlateID-CenterID e.g. TCGA-02-0021-01A-01D-0002-04	No	Set
17	Matched_Norm_Sample_Barcode	BCR Aliquot Barcode for normal sample, e.g. TCGA-02-0021-10A-01D-0002-04 (as opposed to 01A)	No	Set
18	Match_Norm_Seq_Allele1	matched normal sequencing allele or nt (not tested)	No	No
19	Match_Norm_Seq_Allele2	matched normal sequencing allele 2 or nt (not tested)	No	No
20	Tumor_Validation_Allele1	tumor genotyping (validation) allele 1	Yes	No
21	Tumor_Validation_Allele2	tumor genotyping (validation) allele 2	Yes	No
22	Match_Norm_Validation_Allele1	matched normal genotyping (validation) allele 1	Yes	No
23	Match_Norm_Validation_Allele2	matched normal genotyping (validation) allele 2	Yes	No
24	Verification_Status	one of Verified, Wildtype, Unknown	No	Yes
25	Validation_Status	one of Valid, Wildtype, Unknown.	No	Yes
26	Mutation_Status	one of Somatic, Germline, LOH, or Unknown	No	Yes
27	Validation_Method	the assay platform used for the validation call	Yes	No
28	Sequencing_Phase	TCGA Sequencing Phase {1,2,…}	No	Set

Index column indicates the order that the columns are expected. The Null column indicates which MAF columns are allowed to have null values. The Fixed column indicates which MAF columns have specified values: a Fixed value of "No" indicates that there are no specified values for that column; a value of "Yes" indicates that the MAF column requires specific values listed in the Description of Values column; a value of "Set" indicates that the MAF column values come from a specified set of known values (*e.g.* HUGO gene symbols).

Any columns that come after the columns described in Table 7 are optional. Optional columns are not validated by the DCC and can be in any order. The current optional columns are listed in Table 8. NOTE: if you add additional columns that are not listed here, please contact the DCC so that this table can be updated.

Table 8 - Optional mutation annotation format file column headers

MAF Optional Column Header	Description of Values
Treated_Status	is mutation from a treated sample? {Treated, Non-treated}
Hypermutated_Status	is mutation from a hypermutated sample? {Hypermutated, Non-hypermutated}
COSMIC_COMPARISON(ALL_TRANSCRIPTS)	Comparison of mutation to COSMIC database
OMIM_COMPARISON(ALL_TRANSCRIPTS)	Comparison of mutation to OMIM database
Transcript	Transcript used for annotation
CALLED_CLASSIFICATION	Should be the same as Variant_classification
PROT_STRING	Annotation of mutation effect at the protein level
PROT_STRING_SHORT	Annotation of mutation effect at the protein level (short form. for example, frameshift would be fs)
PFAM_DOMAIN	Annotation of the protein domain that mutation resides in

MAF File Checks

The DCC Archive Validator (see "5.3 DCC Archive Validator") checks the integrity of a MAF file. Validation will fail if any of the below are not true for a MAF file (Blue text indicate column header names):

1. Column headers text and order must match SOP (Table 7) exactly
2. Values under column headers listed in the SOP (Table 7) as not null must have values
3. If column headers are listed in the SOP as having *fixed* values (*i.e.* a "Yes" in the "Fixed" column), then the values under those column must come from the enumerated values listed under "Description of Values".
4. If column headers are listed in the SOP as having *set* values (*i.e.* a "Set" in the "Fixed" column), then the values under those column must come from the enumerated values of that domain (*e.g.* HUGO gene symbols).
5. All Allele-based columns must contain "nt" (not tested), - (indel), or a string composed of the following capitalized letters: A, T, G, C.
6. If Validation_Status == "Unknown" then Tumor_Validation_Allele1, Tumor_Validation_Allele2, Match_Norm_Validation_Allele1, Match_Norm_Validation_Allele2 can be null (depending on Validation_Status).
7. If Validation_Status == Valid, then Validated_Tumor_Allele1 and Validated_Tumor_Allele2 must be populated (one of A, C, G, T, and -)
8. Verification_Status and Validation_Status should not conflict (e.g. Wildtype vs Valid).
9. Check allele values against Mutation_Status:

a. If Mutation_Status == "Germline", then
Tumor_Seq_Allele1 == Match_Norm_Seq_Allele1 and Tumor_Seq_Allele2 == Match_Norm_Seq_Allele2.
b. If Mutation_Status == "Somatic" and Validation_Status == "Valid", then
Match_Norm_Validation_Allele1 == Reference_Allele and
Match_Norm_Validation_Allele2 == Reference_Allele and (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) != Reference_Allele
c. If Mutation_Status == "LOH" and Validation_Status==Unknown, then
Tumor_Seq_Allele1 == Tumor_Seq_Allele2 and
Match_Norm_Seq_Allele1 != Match_Norm_Seq_Allele2 and
Tumor_Seq_Allele1 = (Match_Norm_Seq_Allele1 or Match_Norm_Seq_Allele2)
d. If Mutation_Status == "LOH" and Validation_Status==Valid, then
Tumor_Validation_Allele1 == Tumor_Validation_Allele2 and
Match_Norm_Validation_Allele1 != Match_Norm_Validation_Allele2 and
Tumor_Validation_Allele1 == (Match_Norm_Validation_Allele1 or Match_Norm_Validation_Allele2).

10. Check allele values against Validation_status:
    a. If Validation_status == "Wildtype", then
    Tumor_Seq_Allele1=Tumor_Seq_Allele2 and Tumor_Seq_Allele1=Reference_Allele
11. Check that Start_position <= End_position
12. Check for the Start_position and End_position against Variant_Type:
    a. If Variant_Type == "Ins", then
    End_position - Start_position == 1, and
    Reference_Allele == "-", and
    (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) == "-".
    b. If Variant_Type is "Del", then
    Reference_Allele != "-", and
    (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) == "-".
    c. If Variant_Type != "Ins" then
    End_position - Start_position +1 == length(Reference_Allele) and
    (Tumor_Seq_Allele1 or Tumor_Seq_Allele2) == length(Reference_Allele).

Verbose Coverage File

The verbose coverage file (VCF) enables significance analysis of the genes with mutations by providing sequence depth at a mutation locus. The format of a VCF is index and column based, and space-delimited. Indexes are similar to FASTA format in that the index indicator is symbol based. An index line begins with a percent-sign (%) symbol and provides the genome build, chromosome ID, and aliquot barcode (*e.g.* TCGA-02-0001-01C-01W-0359-05; see "3.1 BCR Sample Identifiers - Aliquot Barcodes"). Index positions are described in Table 9. Each line between index lines is column-based delimited using spaces. The first two columns provide a positive increasing range of chromosomal coordinates on the plus (+) strand. The last column provides the sequence coverage for that chromosomal coordinate range. Ranges are chosen by overlapping intervals that have the same depth coverage. Those column positions are described in Table 10. Example data is provided in Figure 5. The suffix of the coverage file should be ".vcf". Since VCFs may be very large, all VCFs should be compressed using gzip inside the archive. The format of the VCF file name should be the same as the TR or MAF files. For example:

- hgsc.bcm.edu_GBM.ABI.1.maf

- hgsc.bcm.edu_GBM.ABI.1.tr
- hgsc.bcm.edu_GBM.ABI.1.vcf.gz

A VCF should accompany every MAF file, but not every MAF file requires a VCF. That is, if a VCF file is submitted, a MAF file must accompany it. However, a MAF file submitted alone does not require a VCF file. VCF files have a data type of "Sequence Coverage" and, by association with MAF files, are considered Level 2 or 3 depending on the MAF file the VCF file accompanies.

Since VCFs have a standardized format, they will be validated eventually. Initially the DCC will spot check VCFs. During one of the DCC's next iterations we will include an automatic check in the DCC Archive Validator. While the validation software is not ready to do the validation yet, it is recommended that you attempt to conform to this specification now. The validation checklist will include:

1. If a VCF file is present in an archive, a MAF file must also be present.
2. Check the order and values of index and column positions listed in Table 9 and Table 10
3. Check that files are space-delimited.
4. The aliquot barcodes listed in the MAF file will be checked to match the aliquot barcodes in the VCF file. An aliquot barcode in a MAF file that cannot be found in the VCF will fail the validation of the archive.

Table 9 - Verbose coverage file index-header descriptions

Index Position	Description
1	The UCSC genome ID (*e.g.* HG18)
2	The chromosome ID (*e.g.* 1, 16, X)
3	The complete BCR aliquot barcode (see "3.1 BCR Sample Identifiers - Aliquot Barcodes")

Table 10 - Verbose coverage file column-header descriptions

Column Position	Description
1	The *starting* position (one-based) of the increasing range of chromosomal coordinates on the plus (+) strand in the chromosome described on the index line
2	The *ending* position (one-based) of the increasing range of chromosomal coordinates on the plus (+) strand in the chromosome described on the index line
3	Indicates the depth of coverage (0, 1, 2, etc reads) for that site-sample. A covered base: >10/15 of the bases surrounding the position of interest are > Phred30

```
%HG18 1 TCGA-13-0799-01A-01D-0359-05
1 200 0
201 240 1
241 260 2
261 278 3
```

Figure 5 - Verbose Coverage File (VCF) example

Data Dependencies

The DCC is dependent on data provided by a GSC, specifically metadata that is meant for deposition at the NCBI Trace or Short Read Archives. The metadata in those archives are used for tracking and reporting. Inconsistencies in those metadata may cause discrepancies in the following report:

- http://tcga-data.nci.nih.gov/tcga/dataSummary.htm